图书情报工作
第62卷 第8期 2018年4月

基于支持向量机的大学生网络信息偶遇影响因素研究*

田梅[1,2] 朱学芳[2]

[1] 新乡医学院管理学院,新乡医学院卫生信息资源研究中心 新乡 453003

[2] 南京大学信息管理学院 南京 210023

摘要:[**目的/意义**]研究网络环境下大学生群体的信息偶遇敏感影响因素,以指导大学生群体提高信息偶遇能力,继而提升大学生信息素养。[**方法/过程**]使用信息增益分析各影响因素与信息偶遇发生频次之间的相关性,构建敏感影响因素模型,并进一步利用支持向量机(SVM)建立信息偶遇频次预测模型。[**结果/结论**]与发生信息偶遇最相关的10个影响因素分布于信息用户、偶遇信息、网络环境、情境因素4个维度;模型分类预测精度达82.96 %,说明SVM对预测信息偶遇频次有良好效果。

关键词:信息偶遇 信息行为 支持向量机 影响因素 信息增益

分类号:G250

DOI:10.13266/j.issn.0252-3116.2018.08.011

1 引言

在信息查询、网络浏览及信息交互的过程中,我们往往会意外收获"感兴趣"或觉得"有用"的信息,这种非目的性偶然获得所需信息的现象就是信息偶遇。虽然信息偶遇是"无预期""意外的"的收获,但信息偶遇不仅可以拓展个体知识面,还可以通过为个体提供更多有用或者感兴趣的信息,无形中促成新的解决问题思路。信息偶遇作为一种被动信息获取的方式,越来越多地受到重视,许多研究已经表明,信息偶遇在个体工作、生活、学习与科研中都发挥着重要作用。在当前Web2.0网络环境下,随着移动互联网的广泛应用,信息的多源与高密度以及用户频繁使用网络及行为"碎片化"等特点,更易激发信息偶遇。大学生群体是使用移动互联网及各种新媒体软件的主要人群之一,对于大学生群体来说,信息偶遇对于获取信息、创新问题解决思路、提高自主学习能力等都有着重要意义。然而不同个体的信息偶遇经验及频次存在很大差别,如何根据大学生群体的信息偶遇特点,结合敏感影响因素,有针对性地研究制定相应策略,从而激发与促进信息偶遇是一个值得思考的问题。

根据文献调研,国内外许多学者围绕信息偶遇的发生以及信息偶遇发生频次进行了相关影响因素研究,从个人因素、信息因素、网络环境等角度提出了研究结果。但是,现有研究中针对大学生群体进行信息偶遇敏感影响因素的研究较少,大多利用访谈法、问卷法、关键事件法、实验法等采集数据并构建影响因素模型,且只是分析各因素的相关性,缺乏相关的定量分析和研究,同时也缺乏对信息偶遇实际应用的研究。而从应用的角度出发,如何有效利用信息偶遇的敏感因素构建相应的决策模型、从而实现对新的未知行为数据的预测分析,是信息偶遇实用化的关键。根据文献调研,目前尚未检索到针对信息偶遇频次进行定量分析的有关文献,同时,关于如何将诸多信息偶遇影响因素应用于未知行为数据的预测也鲜有研究。

作为人工智能的重要分支,机器学习是近年来受到广泛关注的数据分析技术,目前被广泛应用于自然语言处理、计算机视觉等领域。机器学习强调从已有数据中提炼经验和领域知识,并据此改善系统自身性能,最终应用于新的未知数据。若能将实证数据与机器学习相结合,从实际采集的数据出发,构建信息偶遇行为的预测模型,则可以有效促进信息偶遇研究的实际应用效果,并在拓展信息行为研究方法方面做出新的尝试。基于此,本研究将针对大学生群体研究网络

* 本文系2010年国家社会科学基金重大项目"图书、博物、档案数字化服务融合研究"(项目编号:10&ZD134)研究成果之一。

作者简介:田梅(ORCID:0000-0001-6245-8875)副教授,E-mail:tianmeiberry@qq.com;朱学芳(ORCID:0000-0002-8244-5999),教授,博士生导师。

收稿日期:2017-09-15 修回日期:2017-12-30 本文起止页码:84-92 本文责任编辑:易飞

环境下的信息偶遇行为预测问题,引入机器学习中的代表性算法支持向量机(support vector machine, SVM),首先利用信息增益定量分析相应的敏感影响因素,在此基础上,进一步构建针对信息偶遇频次的SVM分类预测模型,并通过采集到的大学生群体信息偶遇调研数据,检验该模型的有效性和合理性。

2 研究回顾与问题提出

在过去的20余年中,诸多学者围绕信息偶遇概念、过程模型及影响因素进行了研究。

2.1 信息偶遇的概念

S. Erdelez于1995年在其博士论文里首次正式提出了信息偶遇(Information Encountering)一词,并把它定义为"在未预期的情境中,个体意外获得感兴趣或可以解决问题的信息的现象。"[1]此后,许多学者提出了相关概念及定义,例如:1996年,K. Williamson等[2]提出了"incidental information acquisition"的概念,将其定义为"在从事其它活动中,出乎意料地获得了信息";学者J. Heinström[3]在此基础上进一步将其定义为"在没有专门查找的情况下,获得有用或有趣的信息";2000年S. Erdelez等[4]提出"information source encountering"的概念,认为在使用网络查找信息时,许多用户在偶遇一个不了解但看起来有用的信息资源时,会有一种机会性获取信息的期望;2000年,E. G. Toms[5]对浏览情境下的信息偶遇进行了探讨与讨论,认为在不同主题间进行信息浏览的过程中,用户会专注于他们意外发现的有趣及有用的信息,并提出了"serendipitous information retrieval"的概念。虽然不同学者的定义及阐述存在差异,但都在表述中强调了信息偶遇过程中用户的"低参与度"与"低预期"两个本质特征以及偶遇信息与用户兴趣或问题相关的突出特点[6-7]。结合国内外学者的相关研究,本文将信息偶遇界定为"是指一种信息获取行为,特指利用网络终端进行各种信息活动时,用户在无目的、低预期的情况下意外获得了自己感兴趣的信息或是觉得有用的信息"。

2.2 信息偶遇过程模型相关研究

M. P. E. Cunha[8]从组织管理角度,提出了一个有助于理解信息偶遇过程的框架模型,模型包括"促成条件(precipitating conditions)""搜寻预设问题A(search for problem A)""双向联想(bisociation)""无预期获得解决问题B的答案(unexpected solution for problem B)"4个部分;L. McCay-Peet等[9]通过对10位历史学者有关"信息搜寻过程"的访谈资料进行分析,在M. P. E. Cunha模型的基础上提出了知识工作中的信息偶遇发生过程模型;V. L. Rubin等[10]对日常生活情境的信息偶遇进行了研究,阐述了信息偶遇发生包括的所有要素方面,并在此基础上,构建了信息偶遇过程要素模型。以上模型研究多侧重于信息偶遇过程中的概念特征要素,也有不少学者从结构化流程的角度对信息偶遇过程模型进行了探讨。S. Erdelez提出了信息搜索情境下信息偶遇发生过程模型,包括注意、停驻、检验、摘取和返回5个功能要素[11];粟村伦久[12]对此模型进行了修订,进一步强调了偶遇信息的利用环节;J. Lawley和P. TOMPKINS[13]提出了基于个体感知的信息偶遇过程模型,将信息偶遇的过程分为6个阶段;S. Makri等[14]基于实证基础,提出了信息偶遇过程模型,强调模型的核心环节是建立"新的某种意识的连接"。

2.3 信息偶遇影响因素研究

S. Erdelez[15]、K. Williamson[2]的研究归纳出了信息偶遇行为的3个基本要素:信息用户、信息环境与偶遇信息。目前有关信息偶遇影响因素的研究多围绕这3个要素展开。

2.3.1 信息用户角度　信息用户的个人特征对于信息偶遇的影响因素研究包括个人特质、信息需求动机、信息素养、信息偶遇经历等方面。J. Heinström[16]认为用户的情绪、个性以及检索风格是信息偶遇的重要影响因素,好奇心强、外向、好学的用户更易发生信息偶遇。台湾学者蔡怡欣等[17]的研究表明,好奇心与求知欲往往激发信息偶遇;个人兴趣和特定的信息需求动机利于激发信息偶遇;经常使用网络,可以轻松处理信息的相关情境,对信息的敏感度相对较高的用户易于发生信息偶遇。田立忠与俞碧飏[18]认为,会把信息困惑放在心里、对检索结果不容易满足、喜欢检索、有广泛浏览习惯和猎奇心理的人,更容易有偶遇的体验;而目的性和策略性很强的人,则不太容易有偶遇经历。郭海霞[19]认为,对信息需求具有内在动机的个体较具有外在动机的个体容易获得信息偶遇经验。袁红与王志鹏[20]在对数字图书馆利用中信息偶遇现象的研究发现,个人因素对信息偶遇的影响大于信息因素,信息偶遇的主观性强,信息用户的信息素养是获取更多偶遇信息的促发剂。S. Erdelez将信息偶遇者分为非偶遇者、巧合偶遇者、偶遇者与超级偶遇者4种类型[1]。其中,超级偶遇者经历信息偶遇的频率非常高,把信息偶遇当作信息搜寻的一种方式,S. Erdelez认为已有的信息偶遇经历是影响信息偶遇发生的重要因素。

2.3.2 息环境角度 信息环境可以理解为信息偶遇发生的不同情境。网络浏览、信息检索、信息交互等都是常见的情境。S. Erdelez[21]研究表明在网络浏览状态下,学者及科研人员易于发生信息偶遇。K. Williamson[2]认为,在与家人、朋友的信息交互中,用户常会经历信息偶遇。潘曙光的研究认为网络信息检索中,用户的前景问题与背景问题可以互换,从而激发信息偶遇;网络浏览中对内容的"低熟悉度""无目标浏览"等因素对信息偶遇的发生有着重要的影响[22]。田立忠与俞碧飏[18]的研究发现在时间压力小、目的性弱且有系统反馈的浏览情境中更易发生信息偶遇。郭海霞[19]认为,在个体没有时间压力下,信息偶遇经验便会越来越多,且无法停止地持续出现。杜雪与刘春茂[23]的研究认为,正在工作状态下的用户更易发生信息偶遇。

2.3.3 偶遇信息角度 S. Erdelez[21,24]将偶遇信息分为两类,即问题相关与兴趣相关,问题相关包括了现在、过去及未来的信息需求。与过去问题相关的偶遇信息,虽然是用户不再需要的信息,不能产生直接的价值,但却能引起用户对信息源的兴趣,进而指导用户未来的信息行为;与现在、将来问题相关的偶遇信息则节省了用户获取信息的时间、精力[24]。V. L. Rubin 等按照用户从信息偶遇中所获得的利益的类型,将用户的信息偶遇结果从"非常抽象的"到"非常具体的"分为3类;其中第二类就描述了用户因为利用与过去、现在问题相关的偶遇信息而获得的利益,即获得关于先前问题或者所关心的事情的解决方案[25]。对于与问题相关的偶遇信息,用户在偶遇发生当时就明确知道其具体用途。和与问题相关的偶遇信息不同,与兴趣相关的偶遇信息多是一些令人感到诧异或惊讶的消息、信息碎片或者可能有用的信息[1],仅供用户娱乐且没有特定用途,但是它能开拓用户的视野、增加用户的知识储备,对用户将来解决问题可能会有帮助,同时它也能帮助用户发现问题[24]。蔡怡欣等[17]认为偶遇信息可以通过自己获得,也可以借由他人获得(直接、间接或公开分享);偶遇信息来源多种多样,包括互联网各类网页、BBS、电子邮件、搜索引擎以及各种网络社交工具与平台等。田立忠与俞碧飏[18]认为信息的外形与位置突出、来源质量高、命名与内容引起用户兴趣、易于获得等信息特点可以提高信息偶遇的概率。郭海霞[19]认为,个人偶遇的信息主要是为满足个人未来可能会有的信息需求做准备,次要是满足个人对于广泛事物的好奇心。

根据以上分析,现有研究成果多为对信息偶遇理论基础的研究,集中在信息偶遇在信息行为框架中的定位和关系分析,缺乏对大学生群体信息偶遇行为的实证研究。其中,尽管有研究[1]分析了信息偶遇频次的重要意义,也未对其进行针对性的定量分析。基于现有成果,结合本研究的对象与目标,提出研究问题:①影响大学生网络信息偶遇的影响因素有哪些? ②诸多影响因素中,哪些因素是影响信息偶遇的敏感因素? ③如何构建的信息偶遇敏感影响因素模型,且预测效果如何?

3 研究设计与数据采集

3.1 研究方法与思路

本研究的目标是从诸多信息偶遇影响因素中,找出敏感影响因素,而关键在于如何从数据角度出发,定量地分析信息偶遇敏感影响因素,并构建分类预测模型。信息增益是度量特征重要程度的有效方法,可表示为一个模型在有和没有一个特征时的信息量的差值,该值越大,意味着该特征能够为预测模型带来的信息量越高,也就意味着该特征越重要、区分度最明显。与开方检验等方法相比,信息增益可从信息论的角度全面地给出特征对于预测模型的重要程度,因此,可有效用于特征选择[26]、属性约简[27]。SVM 是目前具有代表性的机器学习方法。该方法通过需求结构风险最小化,从而实现了在有限样本上良好的泛化能力;通过引入核函数,SVM 提供了非线性问题的有效解决方案[28]。这与本研究的小样本量以及信息偶遇行为在数据分析层面表现出的非线性和交叉性相契合,因此,本研究引入 SVM 和信息增益作为主要数据分析工具,其中,信息增益用来定量分析各影响因素与信息偶遇发生频次的相关关系[29],而 SVM 用来建立有限样本下的信息偶遇频次预测模型。具体研究方法与思路如图1所示,可归纳为以下步骤:

(1)通过对现有相关研究进行整理,荟萃分析整合相关研究结论,从而对现有研究中的信息偶遇影响因素进行收集、整理,并分析影响因素指标,进行归类、区分维度;

(2)参考现有研究中关于信息偶遇影响因素的讨论结果,设计访谈提纲,通过访谈进一步获取更全面的影响因素指标;

(3)依据通过文献与访谈收集而来的影响因素整合结果设计、发放调查问卷;

(4)用信息增益分析各影响因素与信息偶遇发生

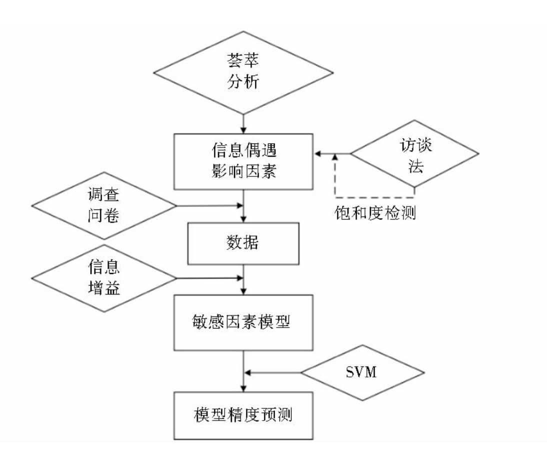


图 1　研究方法与思路

频次之间的相关性,确定敏感因素,构建敏感影响因素模型;

(5)利用 SVM 对测试集数据的信息偶遇频次特点进行预测,并对敏感影响因素模型进行误差预测与分析。

3.2　数据采集

通过文献调研和访谈法获取有关信息偶遇影响因素,并进行统计、整理、归类,最终形成了"信息用户""偶遇信息""信息环境""情境"4 个维度,涵盖"基本信息""人格特质""学习动机"等 12 个方面,共 32 个具体影响因素变量。

3.2.1　信息偶遇影响因素变量

(1)信息用户维度变量。包括信息偶遇用户个人的基本信息(Aa 性别、Ab 专业、Ac 年级)、人格特质(Ad 性格)、学习动机(Ae 好奇心、Af 求知欲)、信息素养(信息意识与信息能力:Ba 信息需求表达、Bb 信息源评估能力、Bc 常用网络工具的使用能力、Bd 熟练使用搜索引擎;检索风格:C 快速检索风格、D 广泛浏览型检索风格、E 深度挖掘型检索风格)、个人经历(X 信息偶遇经历)等 5 个方面的 14 个具体变量。

(2)偶遇信息维度变量。包括偶遇信息的信息热度(F 信息人气、G 信息新颖、H 信息标题感兴趣)、信息质量(I 信息质量权威、科学)、信息内容(J 偶遇信息非常有用、K 偶遇信息与目前问题相关、L 偶遇信息与过去问题相关、M 偶遇信息与将来问题相关、N 偶遇信息是兴趣相关)等 3 个方面的 9 个具体变量。

(3)情境维度变量。包括信息偶遇发生时的信息行为情境(O 信息浏览、P 信息搜索、Q 信息交流)、任务情境(R 任务情境、S 娱乐休闲、V 时间充足)等 2 个方面的 6 个具体变量。

(4)信息环境维度变量。包括信息偶遇发生的网络环境(T 手机上网、U 电脑上网)、系统设计(W 系统反馈)等 2 个方面的 3 个具体变量。

3.2.2　问卷发放与回收

针对上述 32 个具体影响因素变量与最终预测指标"信息偶遇频次"设计调查问卷,共计 33 个问题。经问卷星网站发布问卷,调查对象涉及新乡医学院医学、管理学、心理学 3 个专业 5 个年级的学生,共收回有效问卷 194 份。通过 SPSS22.0 对问卷的信度进行分析,Cronbach's α(克朗巴哈)系数为 0.846,大于 0.8,因此本问卷的信度可以接受。问卷选项采用李克特 5 点量表,调查对象根据自身对题项陈述的赞同程度选择"非常不同意""比较不同意""一般""比较同意""非常同意",分别赋予对应的权值为 1、2、3、4、5。

4　数据整理与分析

将信息用户、偶遇信息、信息环境及情境 4 个维度的 32 个具体因素作为自变量与信息偶遇频次(因变量)进行信息增益计算并分析其相关性,并构建信息偶遇影响因素模型。根据预测目标"信息偶遇频次"值的大小划分为两个组,问卷中选择"比较同意""非常同意"者,即赋值为 4、5 者为一组,表示信息偶遇频次较高;选择"非常不同意""比较不同意""一般"者,即赋值为 1、2、3 者为另一组,表示信息偶遇频次较低。据此,构建样本 194 个,输入样本维数为 32,输出样本维数为 2(频次高和频次低)。首先对上述各个影响因素,计算其与"信息偶遇频次"项的信息增益,根据增益值大小挑选敏感因素,进而引入 SVM 模型,构建信息偶遇预测模型,实现对信息偶遇发生频次的预测与评估。

4.1　计算信息增益

在信息论与概率统计中,熵是表示随机变量的不确定性的度量,而信息增益则可描述具体影响因素带给这些随机变量的信息量。根据定义,首先计算信息偶遇频次的熵,见公式(1):

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \qquad \text{公式(1)}$$

其中 p_i 表示第 i 个随机变量出现的几率,X 为具体影响因素。该熵值越大,表明信息偶遇频次的不确定性越大。其次,计算在已知信息偶遇具体影响因素(即 X)的条件下,信息偶遇频次(即 Y)的条件熵,见公式(2):

$$H(Y|X) = \sum_{i=1}^{n} P(X = x_i) H(Y|X = x_i) \qquad \text{公式(2)}$$

最后,将信息偶遇频次的熵(公式(1))减去信息

偶遇频次的条件熵(公式(2)),即可得到具体影响因素的信息增益,见公式(3):

$$g(Y,X) = H(Y) - H(Y|X) \qquad 公式(3)$$

根据公式(3),可计算得到每个特征的信息增益,根据其值大小,选择信息增益最大的特征为最优特征[29]。

4.2 支持向量机建模

支持向量机是一种适合于小样本的二分类算法。支持向量机建立在统计学习理论基础之上,通过寻求结构风险最小化,可以在有限样本上得到良好的推广能力,避免过学习现象;通过引入核函数,有效解决了高维问题中的"维数灾难"问题,因此在模式识别、故障诊断等领域得到广泛应用[28]。

支持向量机的基本思想是在两类数据之间,寻找一个超平面,使得正负类之间的分类间隔最大。以本研究为例,在"信息偶遇频次低"和"信息偶遇频次高"的数据之间存在多个分类模型,能使得分类间隔最大的 SVM 模型为最终的信息偶遇频次预测模型,如图 2 所示。需要说明的是,为了绘图简便,图 2 只以线性可分问题作为示例,非线性分类和线性不可分问题可以通过 SVM 的核函数和引入惩罚函数进行扩展。

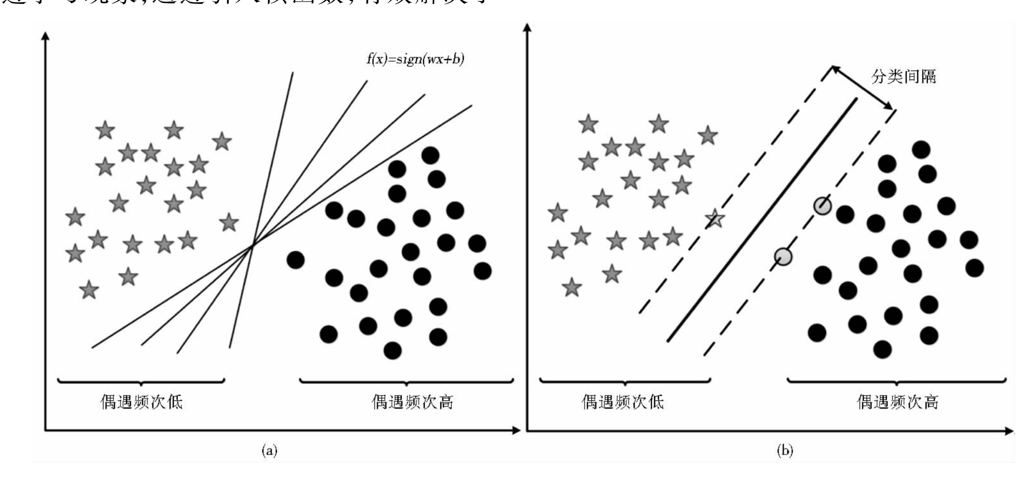


图 2 信息偶遇频次预测模型示意

注:(a)为不同的分类模型,(b)为 SVM 分类模型

图 2 中,以两种不同形状的二维样本点代表不同信息偶遇频次,作为两个不同的类别。从图 2(a)可以看出,可以存在多个分类模型将两类数据完全分开,但哪一个模型为最优模型却无从判断;图 2(b)中,左侧星型点和右侧圆点分别代表代表两类样本,类中间的实线为分类线,两侧虚线分别为穿过距离分类线最近的样本的平行线,它们之间的距离即为分类间隔。根据 SVM 基本原理,能正确区分两类样本、且分类间隔最大的分类线,即为最优分类模型,虚线上的样本点称作支持向量。根据这一原则,通过构建最小化模型,并引入拉格朗日乘子法进行求解,最终可以得到通用的 SVM 分类模型:$f(x) = sign(\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b)$,其中,$\alpha$ 为最优拉格朗日乘子,$K(\cdot)$为处理非线性分类用的核函数,b 为偏置量。SVM 具体理论细节可参考文献[28]。

5 结果与分析

5.1 因变量与各自变量相关性结果

以其信息增益值的大小进行整理排序,结果如表 1 所示:

表 1 信息偶遇频次与各自变量的信息增益

自变量	信息增益值(I)
O 信息浏览	0.241 8
P 信息搜索	0.162 4
T 手机上网	0.141 6
X 信息偶遇经历	0.122 3
N 偶遇信息是兴趣相关	0.113 0
L 信息与过去问题相关	0.111 8
V 时间充足	0.110 0
Bd 熟练使用搜索引擎	0.103 8
J 偶遇信息非常有用	0.093 4
M 信息与将来问题相关	0.093 0
S 娱乐休闲	0.091 3
U 电脑上网	0.090 5
Q 信息交流	0.089 2
Ae 好奇心	0.088 5
E 深度挖掘型检索风格	0.086 9
Bb 信息源评估能力	0.083 8
R 任务情境	0.081 1

（续表 1）

自变量	信息增益值(I)
W 系统反馈	0.076 8
D 广泛浏览型检索风格	0.076 6
C 快速检索风格	0.074 5
Ad 性格	0.061 9
K 信息与目前问题相关	0.057 7
Bc 常用网络工具的使用	0.051 9
Af 求知欲	0.051 6
G 信息新颖	0.041 7
I 信息质量高	0.041 7
Ba 信息需求表达	0.035 9
H 信息标题感兴趣	0.030 4
F 信息人气	0.027 7
Ac 年级	0.029 9

为直观描述，图 3 给出了各属性的信息增益排序。

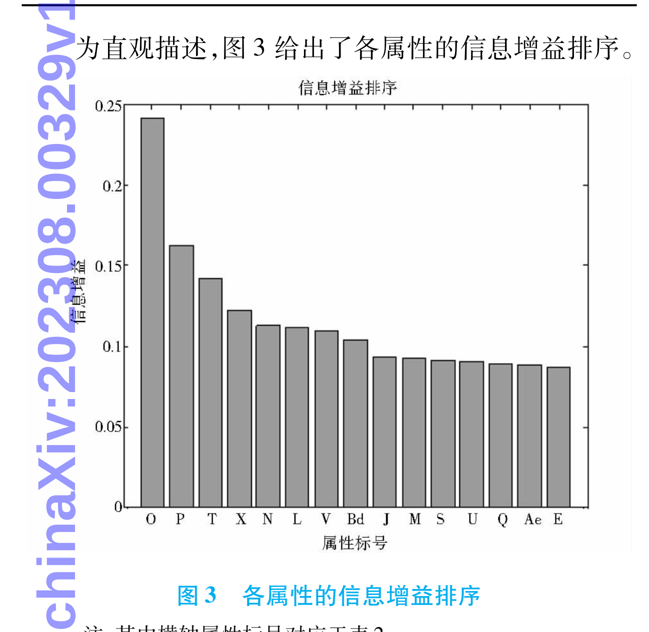


图 3 各属性的信息增益排序

注：其中横轴属性标号对应于表 2

根据公式(4)可知信息增益值 I 应大于等于零，且 I 越大，表明相关性越大。由表 2 和图 3 可知与信息偶遇频次最相关的前 10 个自变量分别是：O 信息浏览、P 信息搜索、T 手机上网、X 信息偶遇经历、N 偶遇信息是兴趣相关、L 信息与过去问题相关、V 时间充足、Bd 熟练使用搜索引擎、J 偶遇信息非常有用、M 信息与将来问题相关。

5.2 模型构建与精度分析

5.2.1 信息偶遇敏感影响因素模型构建　根据上述结果，建立信息偶遇敏感影响因素模型见图 4。

5.2.2 模型精度分析　基于获得的信息偶遇敏感影响因素，本研究使用支持向量机作为分类器，预测信息偶遇频次。首先将获得的 194 个问卷分别提取前 10 个敏感因素，并按照对应的信息偶遇频次高低，构建样本集；其次，采用随机挑选的 135 个样本为训练集，放入支持向量机进行训练，构建一个信息偶遇频次预测模型；最后，采用其余 59 个样本为测试集，放入到该预测模型中进行预测，评价模型预测效果。实验中，按照所有样本进行归一化预处理，SVM 采用 RBF 核函数 $k(x,y)=\exp(-\sigma\|x-y\|^2)$，其中 σ 为核参数。SVM 工具箱采用 LibSVM，使用网格搜索与 5 折交叉验证的方式进行模型选择，确定正则化参数 C 和核参数 σ 的最优组合，参数寻优过程如图 5 所示。预测效果评价采用测试集上的分类正确率，即测试集中预测正确的样本数目在整个测试集中的比例。

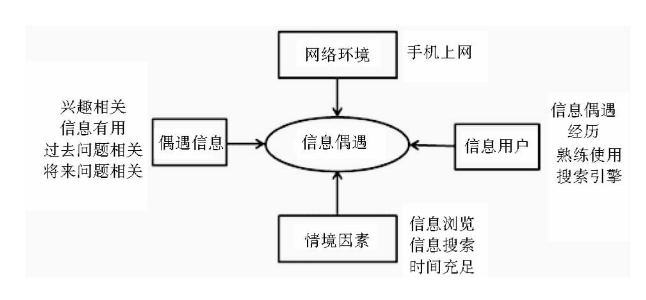


图 4 信息偶遇敏感影响因素模型

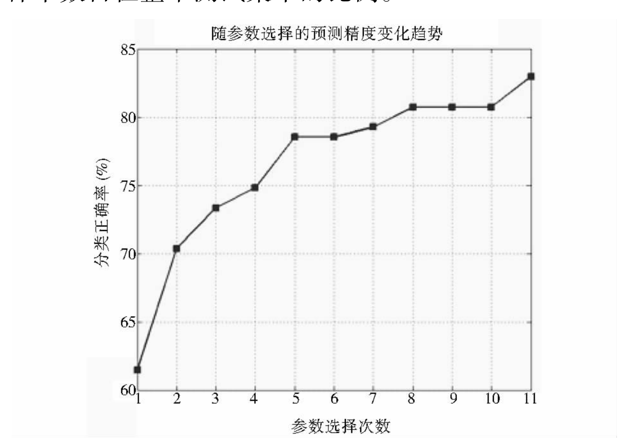


图 5 参数寻优过程

图 5 中，横坐标为参数组合序号，该组合由 LibSVM 自动搜索得到，对应的分类正确率值见表 2。

表 3 显示，通过 SVM 的参数优化选择，最终在测试数据上的分类正确率为 82.96%，取得了较好的预测效果，这表明 SVM 适用于用户信息偶遇行为预测与分析研究，同时，也证明了利用图 2 所找到的 10 个敏感因素的有效性，间接说明了本文所提方法的合理性。但是，从表 3 也可看出，该模型并未取得非常高的分类预测精度，分析原因，可能与调研对象局限于一个高校、行为特点与信息素养存在同质性以及个别问卷题目设计相似度较高等造成数据区分度不高有关。

表2 SVM模型选择参数组合

序号	1	2	3	4	5	6	7	8	9	10	11
C	0.125	0.125	0.125	0.25	0.25	0.25	0.5	0.5	1	2	2
0.062 5	0.125	0.25	0.062 5	0.125	0.25	0.062 5	0.25	0.062 5	0.062 5	1	
分类正确率(%)	61.48	70.37	73.33	74.81	78.52	78.52	79.26	80.74	80.74	80.74	82.96

5.3 结果分析

由上述结果可以看到,本文从多途径尽可能多地提炼信息偶遇影响因素指标,并从构建信息偶遇频次预测模型的角度找到10个最为相关的敏感因素。从图2可以看出,这10个因素与已有研究结果基本保持一致,分布于信息用户、偶遇信息、网络环境、情境因素4个维度。这不仅从另一个角度印证了前期研究,也为有效获取敏感影响因素提供了新的应用方法。

5.3.1 情境因素维度 在10个最相关因素中,情境因素维度中的"O信息浏览"与"P信息搜索"信息增益值居前两位,即信息偶遇多发生于浏览信息时与信息搜索时。根据俞碧飏[30]的观点,一次信息偶遇会经历"信息行为-信息获取-信息需求"的过程,特定信息行为情境中(信息浏览、信息搜索、信息交流)意外获取信息,从而引发新的信息需求。信息浏览情境下,用户处于"目的性较弱",甚至"无目的"的状态,对于结果也是"低预期",甚至"无预期",这种行为特征恰与信息偶遇的特征相契合。此外,信息浏览一般还具有时间压力小的特征[18],根据本研究结果,"V时间充足"也是与发生信息偶遇比较相关的因素之一(I=0.110 0,排第7位),因此,在信息浏览中更易发生信息偶遇。信息搜索往往处于某种任务情境,有较明确的信息需求,时间相对不充足,已有研究多认为此种情境下不易发生信息偶遇。然而,在信息搜索过程中,以超链接方式进行组织的海量的、高密度的网络资源会促使信息偶遇发生,加之目前各种搜索引擎的个性化设置与推荐,无疑会提高信息偶遇发生的几率。需要注意的是,信息时代下大学生对信息的需求呈现快速增加的态势,无论学习、科研、就业及日常行为,均明显依赖于各种外部及互联网信息,尤其是当前移动互联网环境下,移动终端的普及使得大学生获取各类信息更加便捷。信息浏览与信息搜索是大学生群体获取信息的主要途径[31],培养在这两种情境下的信息偶遇能力对于提高大学生的信息素养有着重要意义。

5.3.2 网络环境维度 本研究中,关于网络环境设计了两个问题,分别是"手机上网"与"电脑上网"对信息偶遇发生的影响。其中手机上网代表移动互联网环境,电脑上网代表传统互联网环境。结果显示,"T手机上网"与发生信息偶遇更为相关,并且排在第3位(I=0.141 6)。移动互联网环境下,用户网络使用频次明显增加,根据2012年中国互联网络信息中心(CNNIC)《中国手机网民上网行为特点》的调查[32],72.2%的手机网民每天至少通过手机上网一次,其中,近6成手机网民每天使用手机上网多次。艾瑞咨询统计的《2014年中国移动互联网用户行为研究报告》[33]显示,67%的手机用户表示"每天使用多次"。大学生是移动互联网的主要用户群体之一[34],一方面,在频繁使用网络的过程中,动态新闻、推送信息、交流信息等持续为用户提供最新信息,这本身就可以激发信息偶遇的产生;另一方面,移动互联网下,用户"无目的"的碎片化行为以及信息交流意愿强烈等特征也易于激发信息偶遇。

5.3.3 信息用户维度 研究结果显示,在个人相关的诸多因素中,"X信息偶遇经历""Bd熟练使用搜索引擎"两个变量与发生信息偶遇的相关性较高。其中,"X信息偶遇经历"所指的个人曾经的信息偶遇经历对发生信息偶遇有着较强的影响(I=0.0.122 3,排第4位)。根据A. E. Foster等的研究,信息偶遇不但可以强化个人对问题的理解,修正用户对初始问题的理解,还可以将用户引入一个新的方向,找到解决问题的新思路[35]。这种信息旅程的转向,能产生积极的行动效果与积极的情绪体验[36]。意外获取信息解决问题、找到新的思路对于用户来说是宝贵的信息获取经验与兴奋的情绪体验,积极地影响着下一次信息偶遇的发生。"Bd熟练使用搜索引擎"代表着用户较高的个人信息素养,有着较强的信息意识与信息能力。一般情况下,对信息的敏感度相对较高,经常使用网络、熟练掌握常用工具,可以轻松处理信息相关情境的用户,更易发生信息偶遇[17]。另一方面,熟练使用搜索引擎更易掌握使用技巧,利用个性化推荐等功能获取大量信息,从而激发信息偶遇。与主动信息获取相同,信息偶遇能力也可以进行培养与提高,可以利用现行信息素养实践模型,在实践的各个环节中,引导培养大学生的信息偶遇意识、激发其信息偶遇的能力[37]。

5.3.4 偶遇信息维度 排在前10位的相关因素中,有"N偶遇信息是兴趣相关""L信息与过去问题相关"

“J 偶遇信息非常有用”“M 信息与将来问题相关”4 个变量属于偶遇信息维度。其中，N 变量代表偶遇信息是兴趣相关，J 变量代表偶遇信息是问题相关。根据研究结果，“N 偶遇信息是兴趣相关”对于发生信息偶遇更为相关(I＝0.0.113 0，排第5 位)。兴趣相关的信息偶遇多发生于无任务浏览情境下，问题相关的信息偶遇多发生于有任务搜索情境下[20]。在目前 Web2.0 环境下，大学生群体是各类社交媒体软件的主要用户人群之一，利用碎片化时间频繁使用网络，以“碎片化”行为被动获取着各类平台推送的或者交流而来的“碎片化”信息，这种情况下，被调查的大学生群体认为他们偶遇的信息是“兴趣相关”多于“问题相关”。研究中，与问题相关的因素又分为了“过去问题相关”“目前问题相关”与“将来问题相关”，分别对应着 K 变量、L 变量与 M 变量。结果显示，“过去问题相关”与“将来问题相关”的信息更易激发信息偶遇。其中，“L 信息与过去问题相关”排在第 6 位(I＝0.0.111 8)。在学习与生活中，往往会有一些当时解决不了的问题，而这些问题会作为“背景问题”存储于用户的潜意识中[38]，被动获取的信息与“背景问题”发生关联，即会发生信息偶遇。

6 结语

本文针对大学生群体，围绕“影响大学生网络信息偶遇的影响因素”以及“影响信息偶遇的敏感因素”及“对未知行为数据进行信息偶遇频次特点预测”3 个核心问题进行了实证研究，从机器学习的角度出发，构建了信息敏感偶遇影响因素模型和行为预测模型，并对结果进行理论阐释与分析。

在移动互联网时代，信息偶遇能力是大学生信息素养的重要组成部分。信息偶遇理论研究的最终目的是应用，基于这一认识，本研究将机器学习引入信息偶遇频次预测，是对信息偶遇实用化的有益尝试。一方面，可有针对性地指导大学生提升信息素养，提高自主学习能力；另一方面，本研究的结果可移植到不同人群的信息行为的特点分析和预测中，例如，高校科研工作者的信息偶遇能力对科研工作的提升效果分析等；同时，本研究也能为各类移动终端应用服务商和软件开发商提供创新依据和理论指导，例如，预测用户在使用社交媒体时的信息偶遇行为，可为用户推荐质量更高的推送信息，这有利于进一步改进产品设计，提高服务质量。

参考文献：

[1] ERDELEZ S. Informationencountering: an exploration beyond information seeking[D]. Syracuse: Syracuse University, 1995.

[2] WILLIAMSON K. Discovered by chance: the role of incidental information acquisition in an ecological model of information use[J]. Library & information science research, 1998;20(1):23－40.

[3] HEINSTROM J. Broad exploration or precise specificity: two basic information seeking patterns among students[J]. Journal of the American Society for Information Science and Technology, 2006,57(11):1440－1450.

[4] ERDELEZ S, TOMS EG, RIOUXK, et al. Opportunistic acquisition of information: the new frontier for information user studies[J]. Proceedings of the American Society for Information Science and Technology, 2002,39(1), 521－522.

[5] TOMS E G. Serendipitous information retrieval[EB/OL]. [2018－02－18]. http://www.doc88.com/p-9723328638065.html.

[6] 潘曙光. 信息偶遇研究[D]. 重庆:西南大学,2010.

[7] 张倩,邓小昭. 信息偶遇利用研究文献综述[J]. 图书情报工作,2014,58(20):138－144.

[8] Cunha M P E. Serendipity: why some organizations are luckier than others[EB/OL]. [2018－02－18]. http://citeseerx.ist.psu.edu/viewdoc/download? doi = 10.1.1.540.9982&rep = rep1&type = pdf.

[9] MCCAY-PEET L, TOMS E G. The process of serendipity in knowledge work[M]. New Jersey: ACM, 2010.

[10] RUBIN VL, BURKELL J, QUAN-HAASE A. Facets of serendipity in everyday chance encounters: a grounded theory approach to blog analysis[J]. Information research, 2011,16(3):488.

[11] ERDELEZ S. Towards Understanding Information Encountering on the Web[C]//Proceedings of the 63rd annual meeting of the American Society for Information Science. Medford, NJ: Information Today, Inc., 2000,47:363－371.

[12] 粟村伦久. 情報遭遇に関する利用者行動モデルの再検討:ウェブ上の情報遭遇に対する調査を基として[J]. Library and information science, 2006(55): 47－69.

[13] LAWLEY J, TOMPKINS P. Maximising serendipity: the art of recognising and fostering unexpected potential - Asystemic approach to change[EB/OL]. [2018－02－18]. http://www.cleanlanguage.co.uk/articles/articles/224/1/Maximising - Serendipity/Page1.html.

[14] MAKRI S, BLANDFORD A. Coming across information serendipitously - part 1[J]. Journal of documentation, 2012, 68(5):684－705(22).

[15] Erdelez S. Information Encountering: It's more than just bumping into information[J]. Bulletin of the American Society for Information Science & Technology, 1999, 25(3):26－29.

[16] HEINSTROM J. Psychological factors behind incidental information acquisition[J]. Library & information science research, 2006, 28(4):579－594.

[17] 蔡怡欣,黄元鹤. 线上咨询偶遇经验与个人特征之研究[J]. 图书馆学与咨询科学,2010,36(2):16－34.

[18] 田立忠,俞碧飏.科研人员信息偶遇的影响因素研究[J].情报科学,2013,31(4):69-75.

[19] 郭海霞.网络浏览中的信息偶遇调查和研究[J].情报杂志,2013,32(4):47-50.

[20] 袁红,王志鹏.数字图书馆利用中信息偶遇现象研究[J].图书情报工作,2014,58(17):104-111.

[21] ERDELEZ S. Investigation of information encountering in the controlled research environment[J]. Information and processing management,2004,40(6):1013-1025.

[22] 潘曙光.不同情境下的信息偶遇研究[J].情报探索,2012(8):15-18.

[23] 杜雪,刘春茂. 网络信息偶遇影响因素个性特征的调查实验研究[J].图书情报工作,2015,59(11):119-126.

[24] ERDELEZ S. Information encountering: a conceptual framework for accidental informationdiscovery [C]//Proceedings of an international conference on research in information needs, seeking and use in different contexts. London:Taylor Graham, 1997: 412-421.

[25] RUBIN V L, BURKELL J, QUAN-HAASE A. Everyday serendipity as described in social media[J]. Proceedings of the American Society for Information Science & Technology, 2011, 47(1):1-2.

[26] 刘汝隽, 贾斌,辛阳. 基于信息增益特征选择的网络异常检测模型[J]. 计算机应用, 2016, 36(S2): 49-53.

[27] 贾平, 代建华, 潘云鹤, 等. 一种基于互信息增益率的新属性约简算法[J]. 浙江大学学报(工学版), 2006, 40(6): 1041-1044, 1070.

[28] 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报, 2001, 40(1): 2-10.

[29] 张振海, 李士宁, 李志刚,等. 一类基于信息熵的多标签特征选择算法[J]. 计算机研究与发展, 2013, 50(6): 1177-1184.

[30] 俞碧飏.信息偶遇概念与特点的实证辨析:以科研人员为例[J].情报学报,2012,31(7):759-769.

[31] 周艳玫, 刘东苏, 王衍喜,等. 大学生信息行为调查分析与信息服务对策[J]. 图书情报工作, 2015,59(6):61-67.

[32] CNNIC.中国手机网民上网行为特点[EB/OL].[2017-03-25]. http://www.cnnic.cn/hlwfzyj/hlwxzbg/ydhlwbg/201211/P020121116518463145828.pdf.

[33] 2014年中国移动互联网用户行为研究报告[EB/OL].[2017-03-25]. http://www.useit.com.cn/thread-5941-1-1.html.

[34] CNNIC. 2015年中国青少年上网行为研究报告[EB/OL].[2017-11-25]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/qsnbg/201608/P020160812393489128332.pdf.

[35] FOSTER A E, FORD N J. Serendipity and information seeking: an empirical study [J]. Journal of documentation, 2002,59(3):321-340.

[36] 周佩,黄春燕.信息偶遇研究元人种志分析[J].图书情报工作,2014,58(14):115-120.

[37] 田梅, 朱学芳. 基于现行信息素养模型的大学生信息偶遇能力培养[J]. 图书情报工作, 2015, 59(17):41-46.

[38] 王文韬,谢阳群.信息偶遇模型研究回顾[J].图书情报工作,2014,58(21):130-135.

作者贡献说明:

田梅:确定选题、构思、撰写论文;

朱学芳:指导论文写作。

Study of Network Information Encountering Influence Factors for Undergraduate Group Based on Support Vector Machine

Tian Mei[1,2] Zhu Xuefang[2]

[1] Research Center of Health Information Resources, Management Institute, Xinxiang Medical University, Xinxiang 453003

[2] School of information management, Nanjing University, Nanjing 210023

Abstract: [**Purpose/significance**] In the current Web 2.0 network environment, information encountering is one important method to get information for the undergraduate group. This study is of important significance of improving the ability of information encountering and information literacy for university students. [**Method/process**] Aiming at university students, this paper studies the sensitive influence factors of information encountering in the environment of network. Specifically speaking, this paper uses information gain to analyze the correlation between each influence factor and information encountering frequency, and then builds the model of sensitive influence factor. Furthermore, support vector machine (SVM) is introduced to establish the prediction model for information encountering frequency. [**Result/conclusion**] There exists 10 most sensitive influence factors for information encountering which are located in four dimensions including information user, encountering information, network environment and situation factors. The predicted classification accuracy can reach 82.96%, which demonstrates SVM works well to predict information encountering frequency.

Keywords: information encountering information behavior support vector machine influence factors information gain